

ASSIGNING CUSTOMERS OF MULTIPLE CLASSES TO PARALLEL SERVERS

Ger Koole

*Dept. of Mathematics and Computer Science, University of Leiden
P.O. Box 9512, 2300 RA Leiden, the Netherlands*

ABSTRACT

Customers of K different classes arrive at M heterogeneous exponential servers. Upon arrival a customer may be blocked or sent to one of the free servers, depending on the state of the servers and the class of the customer. There are class-dependent blocking costs. Consider the optimal policy. We prove the following. When a customer is allowed to enter, it should be sent to the fastest free server. Customers of the class with highest blocking costs should never be blocked. We conclude with two monotonicity results.

OPTIMAL POLICIES; CUSTOMER ASSIGNMENT MODELS; LOSS MODELS

Classification: 90B25 primary, 90C40 secondary.

1. INTRODUCTION AND RESULTS

In Sobel [7] customers of K classes arrive according to independent Poisson processes. There are M parallel servers with different service rates. The service times only depend on the server, not on the class of the customer. Blocking is not allowed, unless the system is full. As the results in this paper are not correct, Sobel & Srivastava [8] wrote a revision in which they prove that the probability of having a full system is minimized by the policy that assigns each customer to the fastest free server.

For one class of customers this result is also obtained by Derman et al. [2] and Hordijk & Koole [3]. With the technique used in Hordijk & Koole [3], the results of Sobel & Srivastava [8], including those of the delay model, can also be proven. However, in this paper we prefer to study a more complex model in which blocking is allowed. We also have general arrival streams.

Consider again M exponential servers with decreasing service rates $\mu_1 \geq \dots \geq \mu_M$ and general arrivals in K classes according to a *Markov Arrival Process* (MAP), which is defined as follows.

Definition. (Markov Arrival Process) Let Λ be the finite state space of a Markov process with transition rates λ_{xy} with $x, y \in \Lambda$. When this process moves from x to y with probability q_{xy}^l an arrival of class l occurs, $\sum_{l \in \{1, \dots, K\}} q_{xy}^l \leq 1$ for all $x, y \in \Lambda$.

Without restricting generality we can assume the sum of transition rates in the MAP to be equal to γ in each state. In Asmussen & Koole [1] it is shown that any point process can be approximated arbitrarily close by an MAP. Hence, using a limiting argument, our results are true for a general arrival process. An action has to be chosen in each state (x, i) , $x \in \Lambda$ and $i \in \{0, 1\}^M$. If $i_j = 1$ there is a customer at queue j . We denote the action with $a = (a_1, \dots, a_K)$, a_l denoting the action for class l . For each class l an arriving customer can either be blocked, $a_l = 0$, or sent to one of the free servers, which is denoted by $a_l = j$ if j is a free server. When a customer of class l is blocked, a blocking cost of b_l is incurred, $b_1 \geq \dots \geq b_K \geq 0$. Note that if a class has negative blocking costs it will always be blocked. Costs are discounted with factor $\alpha \leq 1$. Assume, by rescaling time, $\gamma + \mu_1 + \dots + \mu_M = 1$. Consider the following value function.

$$v_{(x,i)}^{n+1} = \sum_y \lambda_{xy} \left(\sum_{l=1}^K q_{xy}^l \min_{a_l} \left\{ \mathbb{I}_{\{a_l=0\}} (b_l + \alpha v_{(y,i)}^n) + \sum_{j \text{ admissible}} \mathbb{I}_{\{a_l=j\}} \alpha v_{(y,i+e_j)}^n \right\} \right. \\ \left. + \left(1 - \sum_l q_{xy}^l \right) \alpha v_{(y,i)}^n \right) + \sum_{j=1}^M \mu_j \alpha v_{(x,i-e_j \vee 0)}^n. \\ v_{(x,i)}^0 = 0.$$

The expression can be rewritten in the standard dynamic programming form, similar to Sobel [7]. Then (for example by Serfozo [6]) the solution of this discrete-time model gives the optimal policy. The expression is rather complex. When the arrivals are Poisson, with rate λ_l for class l , the expression simplifies to:

$$v_i^{n+1} = \sum_{l=1}^K \lambda_l \min_{a_l} \left\{ \mathbb{I}_{\{a_l=0\}} (b_l + \alpha v_i^n) + \sum_{j \text{ admissible}} \mathbb{I}_{\{a_l=j\}} \alpha v_{i+e_j}^n \right\} + \sum_{j=1}^M \mu_j \alpha v_{i-e_j \vee 0}^n.$$

We omitted the state of the MAP as it consists only of 1 state.

When Λ consists of more than one state, the optimal policy will depend on the state of the arrival process. This means that, generally speaking, the optimal policy is untractable. However, using the following lemma, we can obtain some properties of the optimal policy.

Lemma. *The following equations hold for all n :*

$$v_{(x,i)}^n \leq v_{(x,i+e_j)}^n \quad \text{if } i_j = 0. \quad (1.1)$$

$$v_{(x,i+e_{j_1})}^n \leq v_{(x,i+e_{j_2})}^n \quad \text{if } j_1 < j_2 \text{ and } i_{j_1} = i_{j_2} = 0. \quad (1.2)$$

$$\alpha v_{(x,i+e_j)}^n \leq b_1 + \alpha v_{(x,i)}^n \quad \text{if } i_j = 0. \quad (1.3)$$

$$v_{(x,i+e_{j_1})}^n - v_{(x,i)}^n \leq v_{(x,i+e_{j_2}+e_{j_1})}^n - v_{(x,i+e_{j_2})}^n \quad \text{if } j_1 = \min\{j | (i+e_{j_2})_j = 0\} \quad (1.4)$$

and $i_{j_2} = 0$.

The inductive proof is the subject of section 2. Let us consider the consequences of the lemma. Equation (1.1) states a monotonicity result: the less customers, the better. (Because we did not use $b_k \geq 0$ in its proof, it follows from the monotonicity that blocking is always optimal if $b_k < 0$.) When considering assigning an arbitrary customer to one of the free servers, we have to compare $v_{(x,i+e_j)}^n$ for various j . By (1.2) $v_{(x,i+e_j)}^n$ is minimized by j corresponding to the fastest free server. Equation (1.3) is concerned with the assignment of customers with the highest blocking costs. It says that assigning such a customer to an arbitrary server is better than blocking, i.e. a class 1 customer should never be blocked unless the system is full. Equation (1.4) says that when a class k customer is blocked in state (x, i) , i.e. $v_{(x,i+e_{j_1})}^n - b_k - v_{(x,i)}^n \geq 0$, it is also blocked when there are more customers present (and the state of the MAP is the same). On the other hand, when a customer is admitted, it is admitted as well in states with less customers. Another monotonicity property is the following. If $v_{(x,i+e_{j_1})}^n - b_{k_1} - v_{(x,i)}^n \geq 0$, then also $v_{(x,i+e_{j_1})}^n - b_{k_2} - v_{(x,i)}^n \geq 0$, if $k_1 < k_2$. Thus, when blocking is favorable for class k_1 , then blocking is also favorable for class k_2 . Similarly, when customers of a certain class are admitted, then all customer classes with higher blocking costs are admitted as well. This gives the following.

Theorem. *Optimal discounted and average optimal policies exist and have the following properties:*

If a customer is admitted it should be sent to the fastest free server;

Class 1 customers are never blocked, unless the system is full;

If a class l customer is blocked in (x, i_1) , it is blocked in $(x, i_1 + i_2)$;

If a class l customer is admitted in $(x, i_1 + i_2)$, it is admitted in (x, i_1) ;

If a class l customer is blocked in (x, i) , all classes smaller than l are blocked as well in (x, i) ;

If a class l customer is admitted in (x, i) , all classes larger than l are admitted as well in (x, i) .

Proof. As the state and action spaces are finite, the value function converges to the optimal policy. Note that, by adding dummy transitions in the MAP, we can assume the model to be aperiodic. The lemma gives the properties of the optimal policy. \square

Remark 1. In Hordijk & Koole [3] and [4] we generalized the MAP to a Markov Decision Arrival Process (MDAP), by introducing actions in the arrival process. This way we can model certain types of dependencies in the arrival process. If the MDAP is controlled optimally, then (1.1), (1.2) and (1.3) can also be proven, using the induction method. Equation (1.4) does not hold in this case. See Koole [5] for a counterexample.

Remark 2. Equation (1.1), (1.2) and (1.3) can also be shown using coupling arguments. There seems to be no easy coupling argument proving (1.4). Note that sample path arguments cannot be used here, because different policies are optimal for different realizations of the arrival process.

2. PROOF OF THE LEMMA

By induction, assume the lemma holds up to n . We prove the inequalities for the terms on the arrivals in the K classes and the terms on the departures separately. Multiplying with q_{xy}^l , μ_j etc. and summing gives the complete inequalities for $n + 1$. The terms on arrivals are proven by considering the optimal action on the r.h.s., for each y separately, and then finding an action on the l.h.s. for which the inequality holds.

We start with (1.1). Consider an arbitrary customer class l . Let j_1 be the optimal action in state $(y, i + e_j)$. Take the same action in the l.h.s. Then

$$\alpha v_{(y, i + e_{j_1})}^n \leq \alpha v_{(y, i + e_j + e_{j_1})}^n$$

if $j_1 > 0$ and

$$b_l + \alpha v_{(y, i)}^n \leq b_l + \alpha v_{(y, i + e_j)}^n$$

if $j_1 = 0$, both by induction. Concerning departures, we use

$$\alpha v_{(x, i - e_{j_1})}^n \leq \alpha v_{(x, i + e_j - e_{j_1})}^n$$

for a departure from queue j_1 with $i_{j_1} = 1$, and

$$\alpha v_{(x, i)}^n \leq \alpha v_{(x, i)}^n$$

for a departure from queue j . For departures from empty queues, we use

$$\alpha v_{(x,i)}^n \leq \alpha v_{(x,i+e_j)}^n.$$

Combining the inequalities gives $v_{(x,i)}^{n+1} \leq v_{(x,i+e_j)}^{n+1}$.

We continue with (1.2). Consider an arbitrary customer class l . If the optimal action in $(y, i + e_{j_2})$ is not equal to j_1 , take the same action in $(y, i + e_{j_1})$. Induction gives the inequality. If the optimal action in $(y, i + e_{j_2})$ is sending to server j_1 , we take server j_2 in $(y, i + e_{j_1})$. Then we have equality. The term on the MAP without arrivals goes by induction.

Consider the terms corresponding to departures. Terms for departures from queue j with $j \neq j_1, j_2$ are done with induction. With the help of (1.1) we have

$$\mu_{j_1} v_{(x,i)}^n + \mu_{j_2} v_{(x,i+e_{j_1})}^n \leq \mu_{j_2} v_{(x,i)}^n + \mu_{j_1} v_{(x,i+e_{j_1})}^n \leq \mu_{j_2} v_{(x,i)}^n + \mu_{j_1} v_{(x,i+e_{j_2})}^n.$$

Combining the results gives $v_{(x,i+e_{j_1})}^{n+1} \leq v_{(x,i+e_{j_2})}^{n+1}$.

Consider (1.3). Let j_1 be the optimal action, for a certain customer class l , in state (x, i) . If $j_1 \neq j$, then take action j_1 also in $(y, i + e_j)$, giving

$$\alpha^2 v_{(y,i+e_j+e_{j_1})}^n \leq b_1 + \alpha^2 v_{(y,i+e_{j_1})}^n,$$

by induction. If the optimal action is j , block in $(y, i + e_j)$. Then

$$\alpha b_1 + \alpha^2 v_{(y,i+e_j)}^n \leq b_1 + \alpha^2 v_{(y,i+e_j)}^n.$$

If the optimal action is blocking, take also blocking as action in $(y, i + e_j)$. For departures at servers $j_1 \neq j$ we have

$$\alpha v_{(x,i+e_j-e_{j_1} \vee 0)}^n \leq b_1 + \alpha v_{(x,i-e_{j_1} \vee 0)}^n$$

by induction. For server j we have

$$\alpha v_{(x,i)}^n \leq b_1 + \alpha v_{(x,i)}^n.$$

Multiply the four last inequalities by the appropriate terms and sum to obtain (1.3).

Rewrite (1.4):

$$v_{(x,i+e_{j_1})}^{n+1} + v_{(x,i+e_{j_2})}^{n+1} \leq v_{(x,i+e_{j_2}+e_{j_1})}^{n+1} + v_{(x,i)}^{n+1}.$$

In the following table one can see the optimal actions of the r.h.s. in the left columns and the actions establishing the inequalities in the right columns. Define $j^* = \min\{j | (i + e_{j_1} + e_{j_2})_j = 0\}$. Note that $j^* \neq j_1, j_2$. The terms are denoted with their states. By the definition of j_1 , j^* is never optimal in state i .

$i + e_{j_2} + e_{j_1}$	i	$i + e_{j_1}$	$i + e_{j_2}$	
0	0	0	0	induction
0	j_1	0	j_1	equality
0	j_2	j_2	0	equality
j^*	0	0	j^*	twice induction
j^*	j_1	j^*	j_1	induction
j^*	j_2	j_2	j^*	induction

For example, if, for a certain customer class l , rejection is optimal in i , and if sending a customer to queue j^* is optimal in $i + e_{j_1} + e_{j_2}$, the inequality is established by taking rejection in $i + e_{j_1}$ and action j^* in $i + e_{j_2}$, according to the fourth case in the table. Indeed,

$$v_{(y, i + e_{j_1})}^n - v_{(y, i)}^n \leq v_{(y, i + e_{j_1} + e_{j_2})}^n - v_{(y, i + e_{j_2})}^n \leq v_{(y, i + e_{j_1} + e_{j_2} + e_{j^*})}^n - v_{(y, i + e_{j_2} + e_{j^*})}^n$$

by using induction at both steps, giving

$$b_l + \alpha v_{(y, i + e_{j_1})}^n + \alpha v_{(y, i + e_{j_2} + e_{j^*})}^n \leq \alpha v_{(y, i + e_{j_1} + e_{j_2} + e_{j^*})}^n + b_l + \alpha v_{(y, i)}^n.$$

If $i + e_{j_1} + e_{j_2} = e$, only the first three cases have to be considered.

Regarding the departures we have, concerning server j_1 and j_2 :

$$\mu_{j_1} v_{(x, i)}^n + \mu_{j_2} v_{(x, i + e_{j_1})}^n + \mu_{j_1} v_{(x, i + e_{j_2})}^n + \mu_{j_2} v_{(x, i)}^n \leq$$

$$\mu_{j_1} v_{(x, i + e_{j_2})}^n + \mu_{j_2} v_{(x, i + e_{j_1})}^n + \mu_{j_1} v_{(x, i)}^n + \mu_{j_2} v_{(x, i)}^n.$$

The other departure terms follow by induction.

REFERENCES

- [1] S. Asmussen & G.M. Koole (1993). Marked point processes as limits of Markovian arrival streams. *Journal of Applied Probability* **30**: 365–372.
- [2] C. Derman, G.J. Lieberman & S.M. Ross (1980). On the optimal assignment of servers and a repairman. *Journal of Applied Probability* **17**: 577–581.
- [3] A. Hordijk & G.M. Koole (1992). On the assignment of customers to parallel queues. *Probability in the Engineering and Informational Sciences* **6**: 495–511.
- [4] A. Hordijk & G.M. Koole (1993). On the optimality of LEPT and μc rules for parallel processors and dependent arrival processes. *Advances in Applied Probability* **25**: 979–996.
- [5] G.M. Koole (1992). Stochastic scheduling and dynamic programming. Ph.D. thesis, Leiden University.
- [6] R.F. Serfozo (1979). An equivalence between continuous and discrete time Markov decision processes. *Operations Research* **27**: 616–620.
- [7] M.J. Sobel (1990). Throughput maximization in a loss queueing system with heterogeneous servers. *Journal of Applied Probability* **27**: 693–700.
- [8] M.J. Sobel & C. Srivastava (1991). Full-service policy optimality with heterogeneous servers. Working paper.